FoLiA Tools

Release 2.0.0

Maarten van Gompel

Oct 03, 2019

Contents

CHAPTER 1

Introduction

A number of command-line tools are readily available for working with FoLiA, to various ends. The following tools are currently available:

- `foliavalidator` – Tests if documents are valid FoLiA XML. **Always use this to test your documents if you produce your own FoLiA documents!**
- `foliaquery` – Advanced query tool that searches FoLiA documents for a specified pattern, or modifies a document according to the query. Supports FQL (FoLiA Query Language) and CQL (Corpus Query Language).
- `foliaeval` – Evaluation tool, can compute various evaluation metrics for selected annotation types, either against a gold standard reference or as a measure of inter-annotated agreement.
- `folia2txt` – Convert FoLiA XML to plain text (pure text, without any annotations)
- `folia2annotatedtxt` – Like above, but produces output simple token annotations inline, by appending them directly to the word using a specific delimiter.
- `folia2columns` – This conversion tool reads a FoLiA XML document and produces a simple columned output format (including CSV) in which each token appears on one line. Note that only simple token annotations are supported and a lot of FoLiA data can not be intuitively expressed in a simple columned format!
- `folia2html` – Converts a FoLiA document to a semi-interactive HTML document, with limited support for certain token annotations.
- `folia2dcoi` – Convert FoLiA XML to D-Coi XML (only for annotations supported by D-Coi)
- `foliatree` – Outputs the hierarchy of a FoLiA document.
- `foliacat` – Concatenates two or more FoLiA documents.
- `foliamerge` – Merges the annotations of two or more FoLiA documents into one.
- `foliaid` – Assigns IDs to elements in FoLiA documents
- `foliafreqlist` – Output a frequency list on tokenised FoLiA documents.
- `foliatextcontent` – A tool for adding or stripping text redundancy, supports adding offset information.
- `foliaupgrade` – Upgrades a document to the latest FoLiA version.
- `dcoi2folia` – Convert D-Coi XML (a legacy format) to FoLiA XML
- `conllu2folia` – Convert files in the CONLL-U format to FoLiA XML.

- `rst2folia` – Convert ReStructuredText, a lightweight non-intrusive text markup language, to FoLiA, using *docutils <http://docutils.sourceforge.net/>*.
- `alpino2folia` – Convert Alpino-DS XML to FoLiA XML
- `tei2folia` – Convert a subset of TEI to FoLiA.

All of these tools are written in Python 3. More tools are added as time progresses.

CHAPTER 2

Installation

The FoLiA tools are published to the Python Package Index and can be installed effortlessly using `pip`, from the command-line, type:

```
$ pip3 install folia-tools
```

We use `pip3` to ensure we have the Python 3 version. Add `sudo` to install it globally on your system, but we strongly recommend you use virtualenv to make a self-contained Python environment.

If `pip` is not yet available, install it as follows:

On Debian/Ubuntu-based systems:

```
$ sudo apt-get install python3-pip
```

On RedHat-based systems:

```
$ yum install python3-pip
```

On Arch Linux systems:

```
$ pacman -Syu python-pip
```

On Mac OS X and Windows we recommend you install Anaconda or another Python distribution. The FoLiA tools are also included as part of our own LaMachine distribution.

The source code is hosted on github at https://github.com/proycon/foliatools, once downloaded and extracted, it can also be installed using `python3 setup.py install`.

CHAPTER 3

Usage

To obtain help regarding the usage of any of the available FoLiA tools, please pass the `-h` option on the command line to the tool you intend to use. This will provide a summary on available options and usage examples. Most of the tools can run on both a single FoLiA document, as well as a whole directory of documents, allowing also for recursion. The tools generally take one or more file names or directory names as parameters.

CHAPTER 4

Read more

For more generic FoLiA documentation, see https://folia.readthedocs.io